

Data Mining: Lecture 1

A decorative graphic at the bottom of the slide consisting of several overlapping, wavy, horizontal bands. From top to bottom, the bands are light blue, black, dark grey, and light grey with a fine diagonal line pattern. The bands are separated by thin, slightly wavy lines.

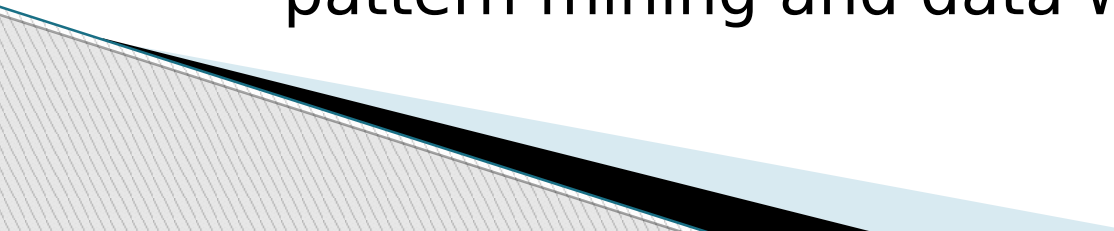
agenda

Course Introduction

Course Details

What is Data Mining?

Course Details

- **Course Description:** The course of Data Mining teaches the students
 - Basic principles, techniques, tools and applications of Data Mining.
 - Science of data mining as the automatic extraction of patterns representing knowledge stored in large databases, data warehouses, and other massive information repositories
 - About the overlap that exists with areas such as machine learning and pattern recognition.
 - The concepts of data pre-processing, cluster analysis, classification and prediction, frequent pattern mining and data warehousing.
- 

Course Resources

- **Text book:**

- Data Mining: Concepts and Techniques (3rd Edition) by Jiawei Han and Micheline Kamber

- **Reference book:**

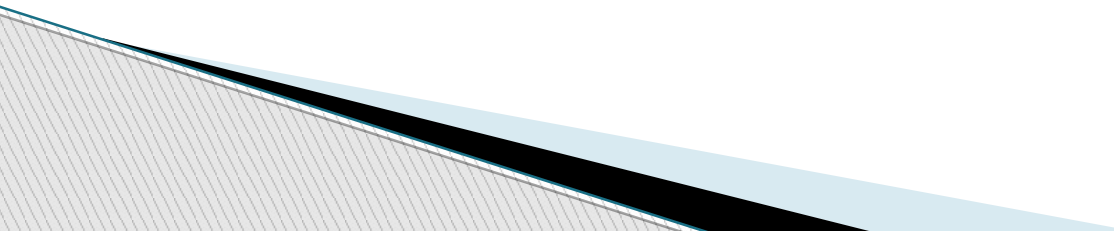
- Elements of Statistical Learning by Hastie, Tibshirani and Friedman
- Freely available online (google for it)

Course Requirement

- You should have some knowledge of the concepts and terminology associated with
 - database systems,
 - statistics,
 - machine learning.

WHAT IS DATA MINING?

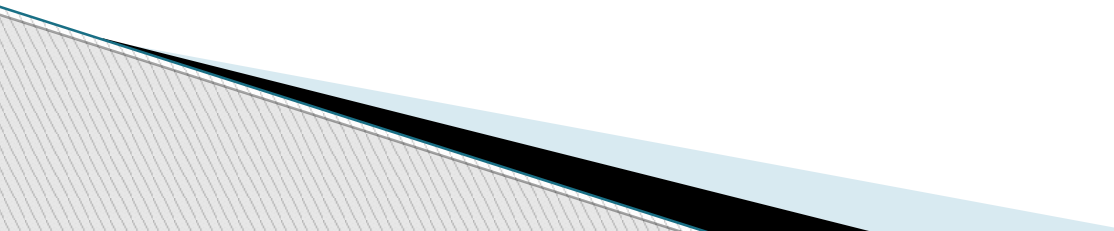
How Data Mining?

- It evolved in to being as the science of databases evolved
 - Database → Datawarehouses → Data Mining
 - Process similar to how science evolved
 - Data Mining and Data Analytics is the fastest growing discipline worldwide with plenty of jobs
- 

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
 - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
 - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
 - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
 - The flood of data from new scientific instruments and simulations
 - The ability to economically store and manage petabytes of data online
 - The Internet and computing Grid that makes all these archives universally accessible
 - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. **Data mining** is a major new challenge!

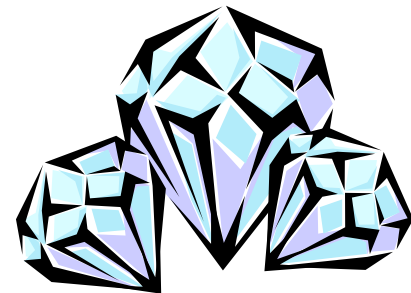
Evolution of Database Technology

- 1960s:
 - Data collection, database creation, IMS and network DBMS
 - 1970s:
 - Relational data model, relational DBMS implementation
 - 1980s:
 - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
 - Application-oriented DBMS (spatial, scientific, engineering, etc.)
 - 1990s:
 - Data mining, data warehousing, multimedia databases, and Web databases
 - 2000s
 - Stream data management and mining
 - Data mining and its applications
 - Web technology (XML, data integration) and global information systems
- 

Data Mining Definition



- Data mining (knowledge discovery from data)
 - Extraction of interesting (implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems



Why Data Mining?

- Huge volumes of Data available: from terabytes to petabytes
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Cameras, publication tools, scanned text and images
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras, YouTube
 - Medical data, demographic data, financial data and marketing data
- We are drowning in data, but starving for knowledge!
- Data mining—Automated analysis of massive data sets

Why Data Mining?—Potential Applications

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
 - Text mining (news group, email, documents) and Web mining
 - Stream data mining
 - Bioinformatics and bio-data analysis

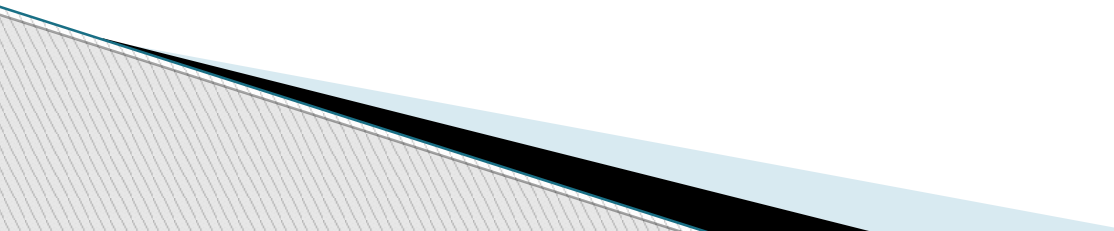
Ex. 1: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
 - Determine customer purchasing patterns over time
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
 - Identify the best products for different groups of customers
 - Predict what factors will attract new customers
- Provision of summary information
 - Multidimensional summary reports
 - Statistical summary information (data central tendency and variation)

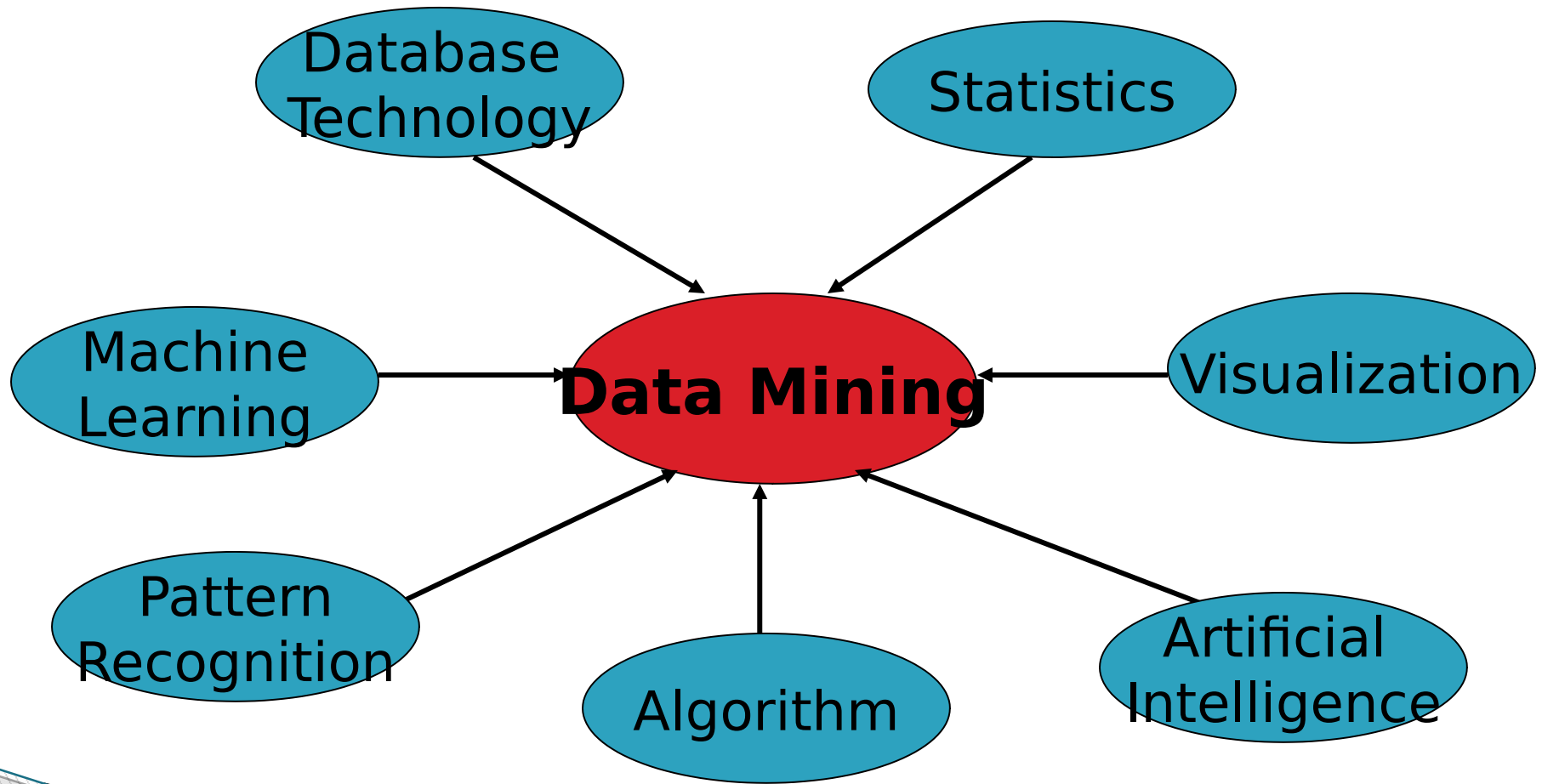
Ex. 2: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
 - Auto insurance: ring of collisions
 - Money laundering: suspicious monetary transactions
 - Medical insurance
 - Professional patients, ring of doctors, and ring of references
 - Unnecessary or correlated screening tests
 - Telecommunications: phone-call fraud
 - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
 - Retail industry
 - Analysts estimate that 38% of retail shrink is due to dishonest employees
 - Anti-terrorism

Ex.4: Biomedical Applications

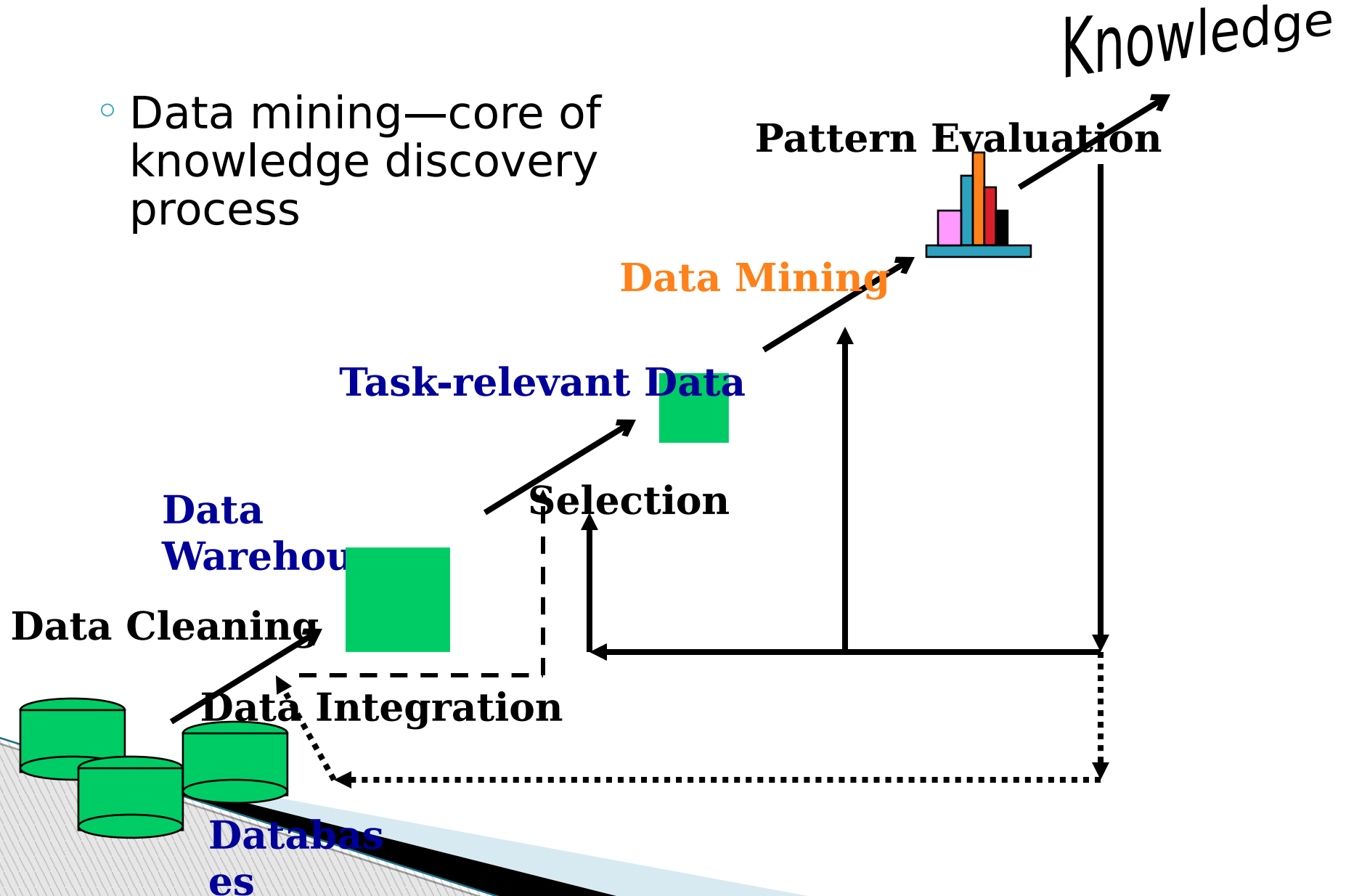
- Approaches: Clustering & Classification
 - Applications:
 - Automated diagnosis
 - Discovery of disease trends
 - Prediction of epidemics
 - Discovering causes for certain conditions
 - Patient data retrieval
- 

Data Mining: Confluence of Multiple Disciplines



Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



KDD Process: Several Key Steps

- Learning the application domain
 - relevant prior knowledge and goals of application
 - Creating a target data set: data selection
 - **Data cleaning** and preprocessing: (may take 60% of effort!)
 - **Data reduction and transformation**
 - Find useful features, dimensionality/variable reduction, invariant representation
 - Choosing functions of data mining
 - summarization, classification, regression, association, clustering
 - Choosing the mining algorithm(s)
 - **Data mining**: search for patterns of interest
 - **Pattern evaluation and knowledge presentation**
 - visualization, transformation, removing redundant patterns, etc.
 - Use of discovered knowledge
- 